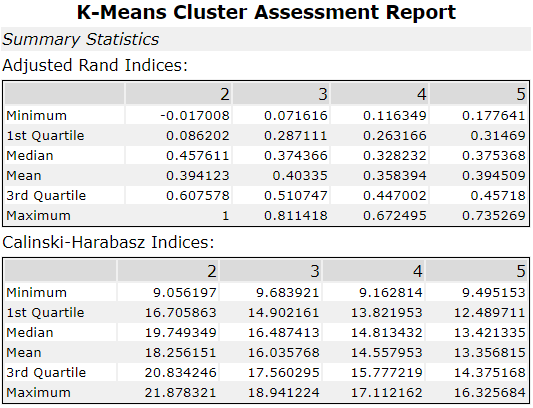
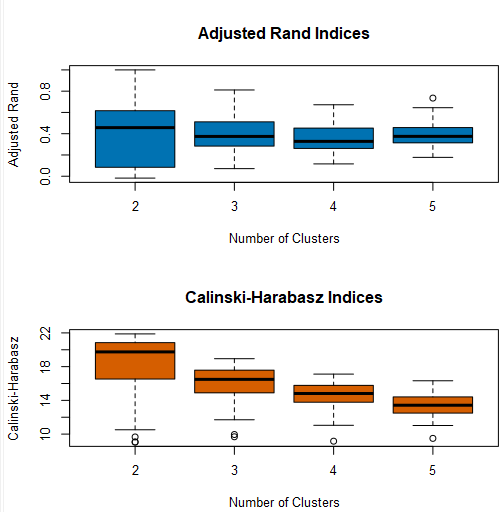
Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?





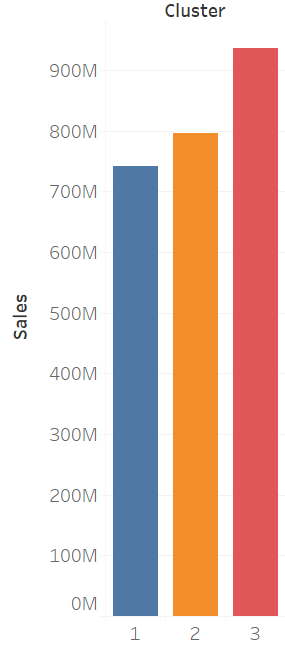
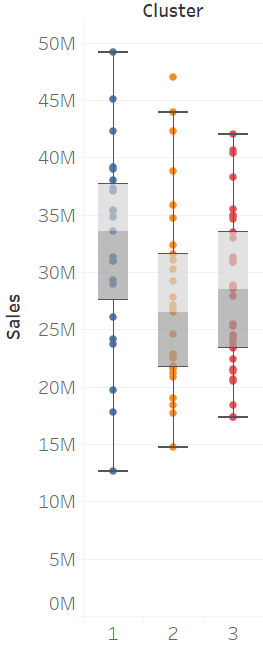
*Analyzing with K-Centroid Diagnostics tool, seems n\_clusters=3 is the best since it has enough mean adjusted rand and tight range.*

1. How many stores fall into each store format?

|  |  |
| --- | --- |
| **Cluster** | **CountDistinctNonNull\_Store** |
| 1 | 23 |
| 2 | 29 |
| 3 | 33 |

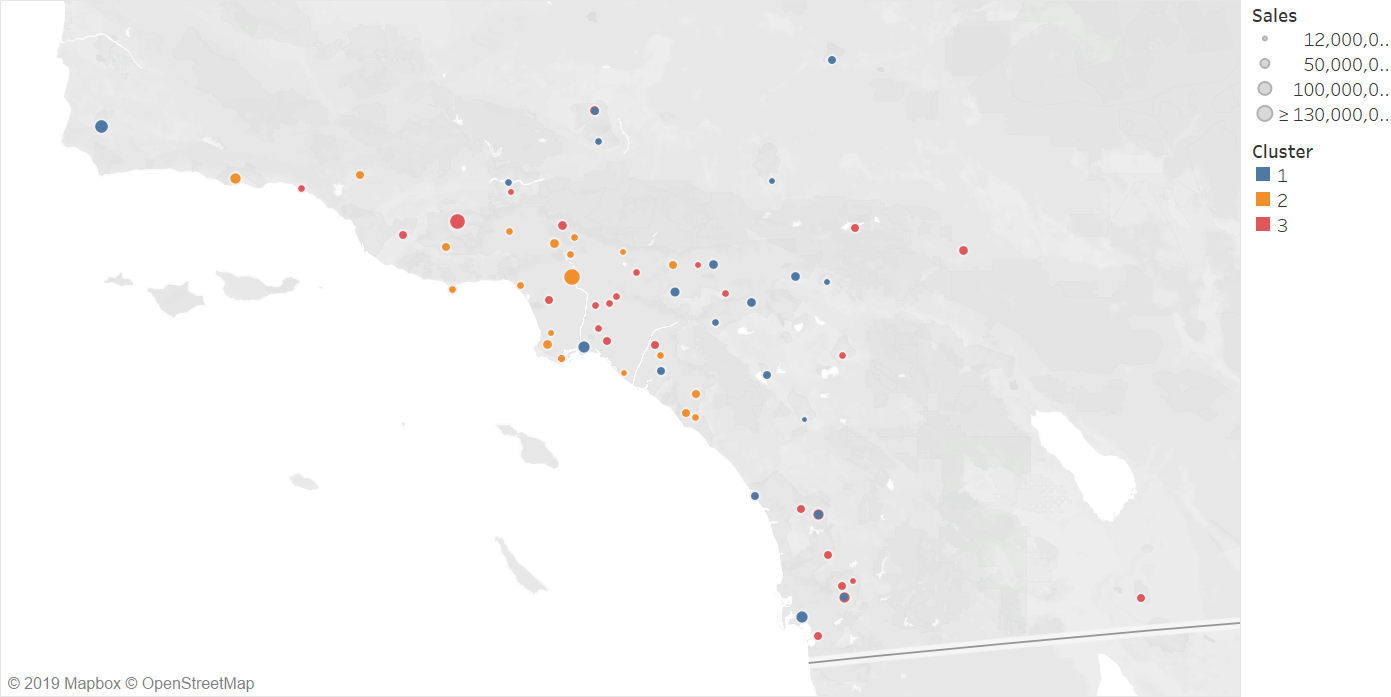
*Using K-centroids cluster analysis, the final result is shown above.*

1. Based on the results of the clustering model, what is one way that the clusters differ from one another?



*Cluster 3 has much tighter sales range, but with higher total sales.*

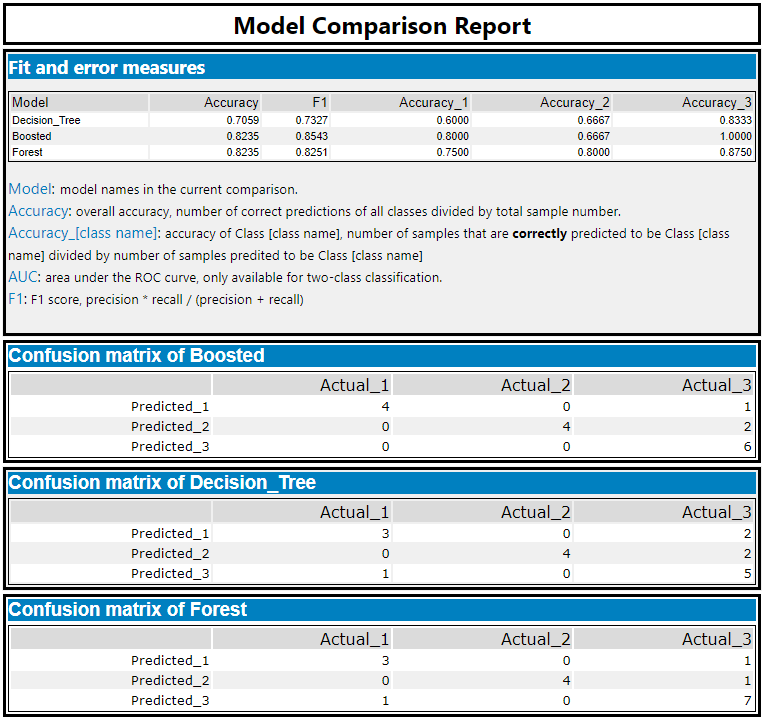
1. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.



## Task 2: Formats for New Stores

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.

*We predict cluster using Decision Tree, Forest Model and Boosted Model. The model comparison report is shown below.*



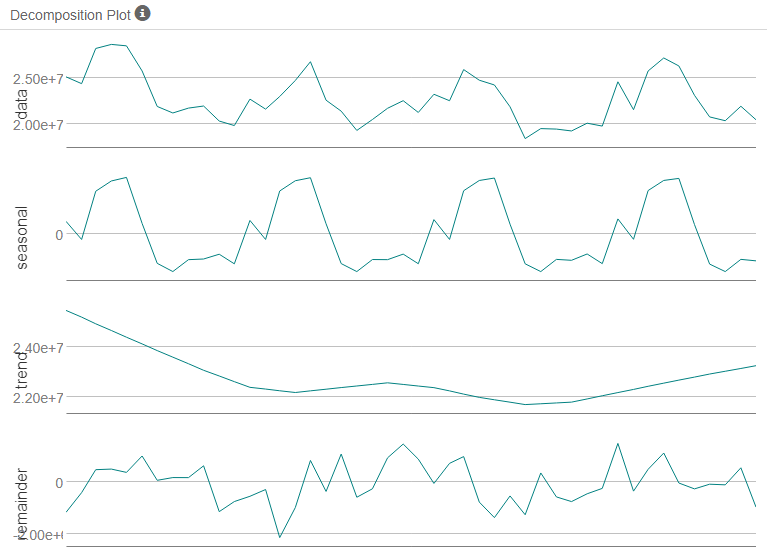
*Judge from the report, the Boosted model has the high accuracy and high F1-score. Thus, we will use the Boosted model for store cluster prediction.*

1. What format do each of the 10 new stores fall into? Please fill in the table below.

|  |  |
| --- | --- |
| ***Store Number*** | ***Segment*** |
| *S0086* | *1* |
| *S0087* | *2* |
| *S0088* | *3* |
| *S0089* | *2* |
| *S0090* | *2* |
| *S0091* | *1* |
| *S0092* | *2* |
| *S0093* | *1* |
| *S0094* | *2* |
| *S0095* | *2* |

## Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?



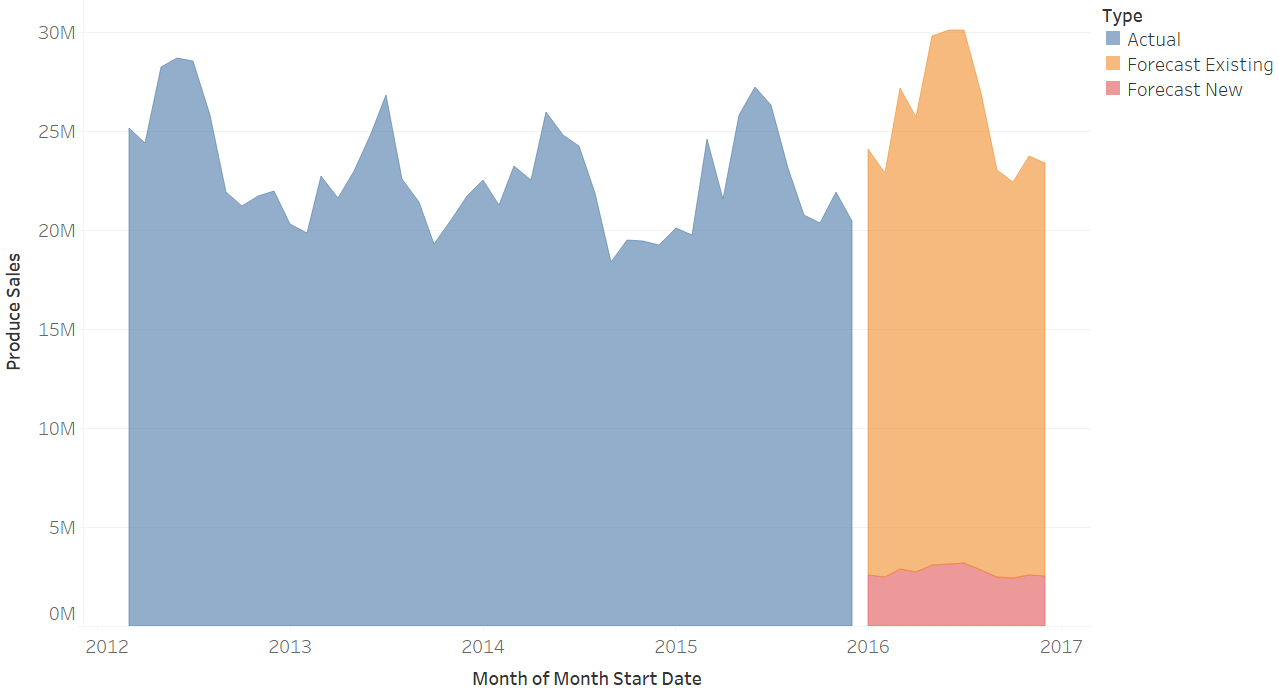
*First, we decomposed the time series plot. We can find that there is no obvious trend, a clear periodic seasonality and increased error element. Thus, the ETS model should be ETS(M,N,M). From the Autocorrelation plot, there is a high serial correlation and seems to be seasonal. The PACF plot also indicates seasonal trend. The ACF plot indicates an exponential decay and sine oscillation. This indicates seasonal difference. Thus we choose ARIMA (0,1,1)(0,1,1) [12] model.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ME** | **RMSE** | **MAE** | **MPE** | **MAPE** | **MASE** |
| ETS | 1983592.6926 | 2226512.5538 | 1983592.6926 | 8.4729 | 8.4729 | 1.2691 |
| ARIMA | 2878344.1382 | 3061362.1418 | 2878344.1382 | 12.5815 | 12.5815 | 1.8416 |

*Judge from the comparison table above, it can be observed that ETS model has lower ME, RMSE, MASE, MPE, MAPE and MASE values. Consequently, ETS model is chosen for forecasting.*

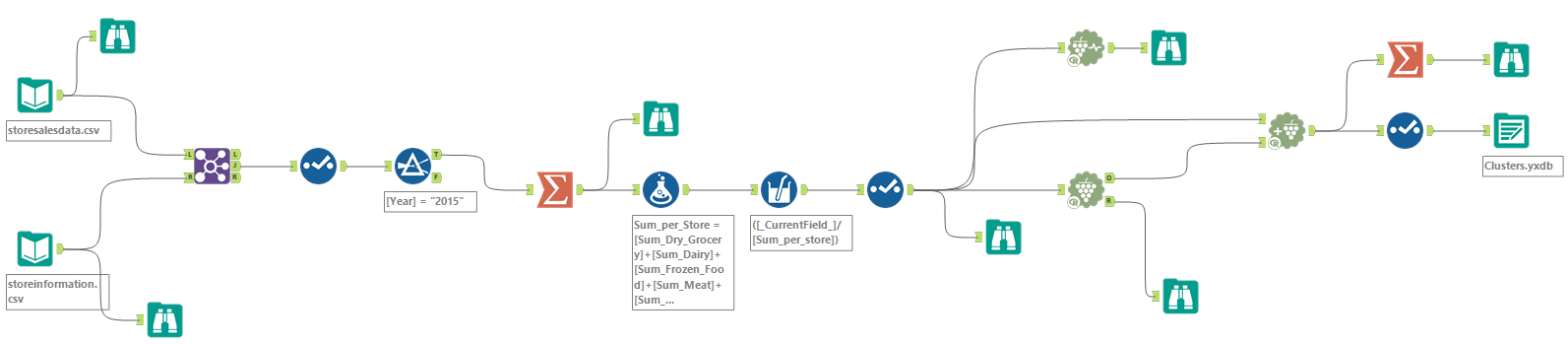
1. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Month** | **New Stores** | **Existing Stores** |
| 2016 | 1 | 2580411 | 21539936 |
| 2016 | 2 | 2494753 | 20413771 |
| 2016 | 3 | 2876480 | 24325953 |
| 2016 | 4 | 2742890 | 22993466 |
| 2016 | 5 | 3103562 | 26691951 |
| 2016 | 6 | 3124176 | 26989964 |
| 2016 | 7 | 3168777 | 26948631 |
| 2016 | 8 | 2820029 | 24091579 |
| 2016 | 9 | 2491912 | 20523492 |
| 2016 | 10 | 2442136 | 20011749 |
| 2016 | 11 | 2551509 | 21177435 |
| 2016 | 12 | 2520758 | 20855799 |

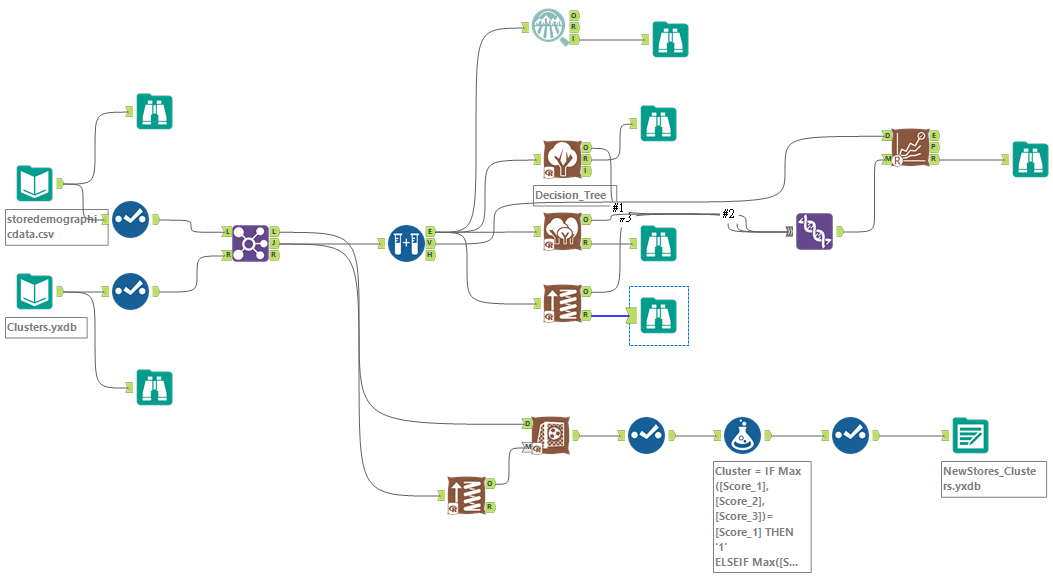


## Appendix: Alteryx Workflows

### Task 1



### Task 2



### Task 3

